Biol. Cybern. 75, 49–57 (1996)

Biological Cybernetics

Transition times in self-organizing maps

Tom M. Heskes

Department of Medical Physics and Biophysics, University of Nijmegen, Geert Grooteplein 21, 6525 EZ Nijmegen, The Netherlands

Received: 23 January 1993/Accepted in revised form: 19 March 1996

Abstract. We study the creation of topological maps. It is well known that topological defects, like kinks in one-dimensional maps or twists ('butterflies') in two-dimensional maps, can be (metastable) fixed points of the learning process. We are interested in transition times from these disordered configurations to the perfectly ordered configurations, i.e., the average time it takes to remove a kink or to unfold a twist. For this study we consider a self-organizing learning rule which is equivalent to the Kohonen learning rule, except for the determination of the 'winning' unit. The advantage of this particular learning rule is that it can be derived from an error potential. The existence of an error potential facilitates a global description of the learning process. Mappings in one and two dimensions are used as examples. For small lateral-interaction strength, topological defects correspond to local minima of the error potential, whereas global minima are perfectly ordered configurations. Theoretical results on the transition times from the local to the global minima of the error potential are compared with computer simulations of the learning rule.

1 Introduction

Sensory signals provide the input to the central nervous system. These signals are represented in sensory maps, which are a crucial first step in the information processing of the brain. The external information is represented in an orderly, topology-preserving manner, i.e., neighboring units in the sensory map code similar input signals. The formation of these maps is a process of self-organization for which several learning paradigms have been suggested (e.g., Von der Malsburg 1973; Takeuchi and Amari 1979; Miller et al. 1989; Durbin and Mitchison 1990). The proposal of Kohonen (1982) does not aim at the modeling of all biological details, but tries to capture the most important features of self-organizing processes. The Kohonen learning rule also has applications in robotics, data segmentation, and classification tasks.

Basically, the algorithm proposed by Kohonen works as follows. Given a certain input vector from the environment, the unit with the smallest Euclidian distance to this vector is called the 'winner'. The weight vector of this unit and, to some extent, its neighboring units, are moved toward the input vector. The properties of this learning procedure, and of closely related variants, have been studied in great detail (Cottrell and Fort 1986; Ritter and Schulten 1986, 1988; Obermayer et al. 1990, 1992).

Recently, much effort has been devoted to the search for an energy function that is minimized by the learning rule (Tolat 1990; Kohonen 1991; Erwin et al. 1992). The existence of such an energy function or error potential facilitates a global description of the performance of the learning procedure. The best possible network state corresponds to a global minimum of the error potential and undesired fixed points of the learning process are simply local minima. Examples of these undesired fixed points are topological defects such as kinks in one-dimensional maps and twists in two-dimensional maps (e.g., Geszti 1990).

In Sect. 2 we will define an error potential for the self-organization of topological maps. This error potential is equivalent to the energy function proposed independently by Luttrell (1994). The corresponding learning rule will be used to study two special examples of topological defects in detail: kinks in Sect. 3 and twists in Sect. 4. In both cases these disordered configurations are true local minima of the error potential. This is illustrated by pictures of the error potentials. We are interested in the transition times from the local minima of the error potential to the global minima, i.e., the average time it takes to remove a kink or to unfold a twist. In Sect. 5 we will review a general study on learning in neural networks with local minima (Heskes et al. 1992). We will apply this general theory to the two specific examples. These theoretical results will be compared with computer simulations of the learning rule in Sect. 6. In Sect. 7 we will discuss the main results.

2 Kohonen learning as the gradient of an error potential

We will use the following definitions: m-dimensional input vectors $\vec{x}$ are drawn from the environment Ω according to a probability density function $\rho(\vec{x})$.

An average with respect to input vectors is denoted by $\langle\,\rangle_\Omega$. The topological map consists of n units, labeled $1, \ldots, i, \ldots, n$. To each unit we ascribe an m-dimensional weight vector $\vec{w}_i$. The combination of all weight vectors is the N-dimensional state vector $\mathbf{w} = (\vec{w}_1^T, \ldots, \vec{w}_i^T, \ldots, \vec{w}_n^T)^T$, so $N = n \times m$. The 'local error' $e_i(\mathbf{w}, \vec{x})$ of a unit i is defined by

$$e_i(\mathbf{w}, \vec{x}) \stackrel{\text{def}}{=} \frac{1}{2} \sum_{j=1}^{n} h_{ij} \|\vec{w}_j - \vec{x}\|^2$$

h is the lateral-interaction matrix with nonnegative elements h_{ij}, independent of the state vector $\mathbf{w}$ and the input $\vec{x}$; usually h_{ij} is a decreasing function of the (physical) distance between unit i and unit j on the topological map. We normalize this matrix by requiring that

$$\sum_{j=1}^{n} h_{ij} = 1 \;\forall_i \tag{1}$$

The 'partition function' $Z_\beta(\mathbf{w}, \vec{x})$ is defined by

$$Z_\beta(\mathbf{w}, \vec{x}) = \sum_{i=1}^{n} \exp[-\beta e_i(\mathbf{w}, \vec{x})] \tag{2}$$

Rose et al. (1990) use this kind of partition function with finite β to describe statistical mechanics and phase transitions in clustering. We use it here to arrive at the usual 'winner-take-all' mechanism in the limit $\beta \to \infty$. This description would appear to be very useful for integration and differentiation.

The global error potential is now the 'free energy' $-\frac{1}{\beta} \ln Z_\beta(\mathbf{w}, \vec{x})$, averaged over the environment Ω, in the limit $\beta \to \infty$:

$$E(\mathbf{w}) \stackrel{\text{def}}{=} -\lim_{\beta \to \infty} \frac{1}{\beta} \langle \ln Z_\beta(\mathbf{w}, \vec{x}) \rangle_\Omega = \langle e_{\kappa(\mathbf{w}, \vec{x})}(\mathbf{w}, \vec{x}) \rangle_\Omega \tag{3}$$

where $\kappa(\mathbf{w}, \vec{x})$ is called the 'winner', the unit with the smallest local error, given the network state $\mathbf{w}$ and the input $\vec{x}$. Note that in the limit $\beta \to \infty$ only the term with the smallest local error survives in the sum in (2). The error potential is thus the smallest error potential averaged over the set of all input vectors. Luttrell (1989) introduced a similar error potential and gave an interpretation in terms of noisy transmission between (neural) layers (see also Ritter et al. 1991). In our terms it can be written (Erwin et al. 1992)

$$E_{\text{Luttrell}}(\mathbf{w}) = \langle e_{\kappa'(\mathbf{w}, \vec{x})}(\mathbf{w}, \vec{x}) \rangle_\Omega \tag{4}$$

where $\kappa'(\mathbf{w}, \vec{x})$ is the unit κ' with the smallest Euclidian distance $\|\mathbf{w}_{\kappa'} - \vec{x}\|$, which most of the time, but not always, will be equal to the $\kappa(\mathbf{w}, \vec{x})$, the unit κ with the smallest local error $e_\kappa(\mathbf{w}, \vec{x})$. Kohonen (1991) derived an *approximate* learning rule starting from the error potential (4). In his recent paper, Luttrell (1994) arrived independently at the same error potential (3).

We will take the error potential (3) as our starting point. The gradient of this error with respect to the state vector $\mathbf{w}$, denoted by ∇, yields

$$\mathbf{f}(\mathbf{w}) \stackrel{\text{def}}{=} -\nabla E(\mathbf{w})$$

$$= \lim_{\beta \to \infty} \left\langle \sum_{i=1}^{n} \frac{-\nabla e_i(\mathbf{w}, \vec{x}) \exp[-\beta e_i(\mathbf{w}, \vec{x})]}{Z_\beta(\mathbf{w}, \vec{x})} \right\rangle_\Omega$$

As in (3), the term with the smallest local error dominates in the limit $\beta \to \infty$. So, the learning procedure that performs *stochastic* gradient descent on the global error potential $E(\mathbf{w})$ defined in (3), is a succession of the following steps:

1. Pick an input vector $\vec{x}$ from the environment Ω according to the probability density function $\rho(\vec{x})$.
2. Find the 'winning unit' κ, i.e., the unit with the smallest local error $e_\kappa(\mathbf{w}, \vec{x})$.
3. Update the weights with

$$\Delta w_{i\alpha} \stackrel{\text{def}}{=} \eta f_{i\alpha}(\mathbf{w}, \vec{x}) = -\eta \frac{\partial e_\kappa(\mathbf{w}, \vec{x})}{\partial w_{i\alpha}} = \eta h_{\kappa i}(x_\alpha - w_{i\alpha}) \tag{5}$$

with η the learning parameter and $\mathbf{f}(\mathbf{w}, \vec{x})$ the so-called stochastic force, a vector with components $f_{i\alpha}(\mathbf{w}, \vec{x})$.

The resulting learning procedure is almost equal to the original learning procedure proposed by Kohonen (1982). As in the discussion above on the distinction between the error potentials (3) and (4), the only difference is the determination of the winning unit, i.e., the second step in the learning procedure. In Kohonen's learning rule the winning unit is the unit for which the Euclidian distance between the weight of that unit and the input vector is the smallest. In our case the winner is the unit with the smallest local error, the *same* error that must be differentiated in order to obtain the learning rule (5). In another paper (Heskes and Kappen 1993; see also Luttrell 1994) we proved that this is the only way to insure that such a winner-take-all learning rule can be derived from a global error potential. Note that in the limit of no lateral interaction, i.e., $h_{ij} = \delta_{ij}$, the learning rule (5) is completely equivalent to the Kohonen learning rule. For nonzero lateral interaction the use of the local errors $e_i(\mathbf{w}, \vec{x})$ for the determination of the winning units leads to a slightly different tesselation of the input space. Qualitatively, properties of the learning procedure defined above are equivalent to properties of the Kohonen learning rule (except for the ordering of a one-dimensional mapping: see Sect. 6); quantitatively there may be some differences.

The existence of a global error potential facilitates a global description of the learning process. The lower the error potential $E(\mathbf{w})$, the better the network state $\mathbf{w}$. As we will see, stable disordered configurations, such as kinks in one-dimensional maps or twists ('butterflies') in two-dimensional maps, are simply local minima of the error potential. The global minima correspond to perfectly ordered configurations (see Bauer and Pawelzik 1992, for a well-defined measure of the preservation or violation of neighborhood relations; in the small examples used in this paper the difference between ordered and disordered configurations will be quite clear). We are

interested in the transition times between the local minima and the global minima. For example, how long does it take (on the average) to remove a kink in a one-dimensional map or to unfold a butterfly in a two-dimensional map?

3 Kinks in a one-dimensional map

We consider a one-dimensional map consisting of three units. The network state vector is written $\mathbf{w} = (w_1, w_2, w_3)^T$. The input x is drawn with equal probability from the interval [0, 1], i.e.,

$$\rho(x) = \theta(x)\theta(1-x)$$

The lateral-interaction matrix h with components h_{ij} is chosen

$$h = \frac{1}{1+\sigma}\begin{pmatrix} 1 & \sigma & 0 \\ \sigma & 1-\sigma & \sigma \\ 0 & \sigma & 1 \end{pmatrix}$$

σ gives the interaction strength between neighboring units in the map; $\sigma = 0$ means no lateral interaction. We will always work with $0 \leqslant \sigma < 1/2$.

Ordered configurations are called 'lines'. One of them, denoted by (123) since $w_1 < w_2 < w_3$, is drawn schematically in Fig. 1a. The other one is (321), i.e., w_1 and w_3 are interchanged. There are four different disordered configurations, called 'kinks': (132), (213), (231) and (312). The first one is sketched in Fig. 1b. By numerical calculations it can be shown that for $\sigma < \sigma^* \approx 0.0822$ the error potential (3) has six minima: two lines are the global minima, four kinks are local minima. At $\sigma = \sigma^*$ the local minima disappear and only two global minima remain.

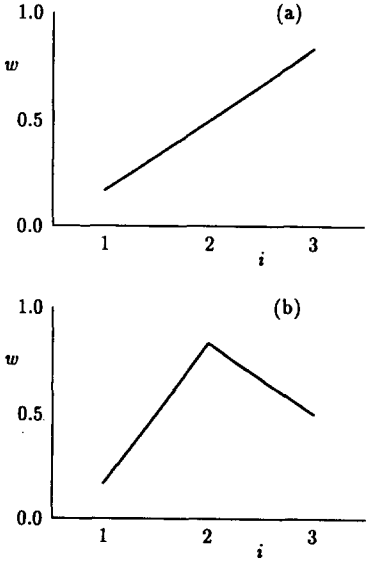

Fig. 1. Configurations in a one-dimensional map: **a** line, **b** kink

We would like to picture how the error potential changes on the way from the local to the global minimum. To remove two degrees of freedom, we define a one-dimensional path through the three-dimensional weight space, subject to the constraints

$$w_1 + w_2 + w_3 = \tfrac{3}{2}$$

and

$$(w_2 - w_1)^2 + (w_3 - w_2)^2 + (w_3 - w_2)^2 = \tfrac{2}{3}$$

Now $\mathbf{w}$ is totally parametrized by the parameter ω:

$$\mathbf{w} = \frac{1}{2}\begin{pmatrix}1\\1\\1\end{pmatrix} + \frac{\sqrt{3}\cos(2\pi\omega)}{9}\begin{pmatrix}-2\\1\\1\end{pmatrix} + \frac{\sin(2\pi\omega)}{3}\begin{pmatrix}0\\-1\\1\end{pmatrix}$$

Since at the minima one of the weights is approximately equal to 1/6, the second to 1/2, and the third to 5/6, the path characterized by ω passes through all the minima. The kinks and the lines are positioned as follows.

Minimum	line	kink	kink	line	kink	kink
Configuration	(123)	(213)	(231)	(321)	(312)	(132)
ω	1/12	1/4	5/12	7/12	3/4	11/12

The error potential as a function of ω is plotted in Fig. 2 for four different values of σ. For $\sigma = 0$ all minima are equally deep (Fig. 2a). For $\sigma > 0$ the kinks have a higher error potential than the lines (Fig. 2b, c). Eventually, the local minima disappear (Fig. 2d).

4 Twists in a two-dimensional map

As a second example, we will study a two-dimensional map consisting of four units. The eight-dimensional state vector reads $\mathbf{w} = (\vec{w}_1^T, \vec{w}_2^T, \vec{w}_3^T, \vec{w}_4^T)^T = (w_{11}, w_{12}, w_{21}, w_{22}, w_{31}, w_{32}, w_{41}, w_{42})^T$. The input $\vec{x} = (x_1, x_2)^T$ is drawn with equal probability from the square $[-1, 1] \times [-1, 1]$, i.e.,

$$\rho(x_1, x_2) = \theta(1+x_1)\theta(1-x_1)\theta(1+x_2)\theta(1-x_2)/4$$

We choose a lateral-interaction matrix h of the form

$$h = \frac{1}{(1+\sigma)^2}\begin{pmatrix} 1 & \sigma & \sigma^2 & \sigma \\ \sigma & 1 & \sigma & \sigma^2 \\ \sigma^2 & \sigma & 1 & \sigma \\ \sigma & \sigma^2 & \sigma & 1 \end{pmatrix}$$

Again σ gives the lateral-interaction strength. We will keep $0 \leqslant \sigma < 1$.

We expect to find possible (local) minima if each unit covers one quadrant of the input space. We denote a particular minimum by $(ijkl)$ if unit i lies in the first quadrant, unit j in the second, and so on. There are $4! = 24$ different possible minima. Eight of them are perfectly ordered. We will call these configurations 'rectangles'. An example of such a rectangle, (1234), is given in Fig. 3a. As usual (Kohonen 1982), lines are drawn between neighboring units on the map, i.e., between 1–2, 2–3, 3–4 and 4–1. For small σ, disordered configurations are local minima of

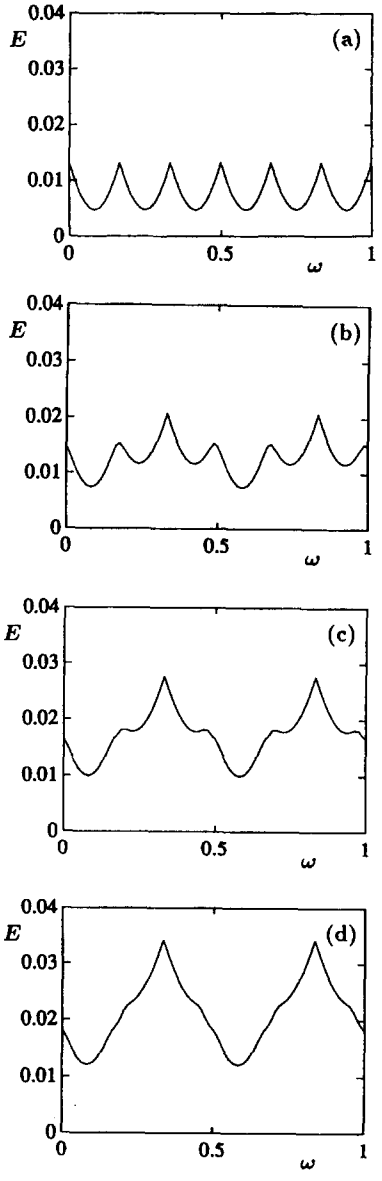

Fig. 2. The error potential E as a function of ω for different values of the interaction strength σ: **a** $\sigma = 0$, **b** $\sigma = 0.04$, **c** $\sigma = 0.08$, **d** $\sigma = 0.12$

the error potential. These minima are called 'twists' or 'butterflies'. In Fig. 3b the twist (1324) is sketched. At $\sigma = \sigma^* \approx 0.240$ the local minima disappear.

To make a picture of the error potential $E(\mathbf{w})$ for this two-dimensional mapping, we have to remove six degrees of freedom. We define a two-dimensional manifold as follows:

1. The four weight vectors $\vec{w}_i$ lie on a circle with radius $r(\sigma)$:

$$\vec{w}_i = r(\sigma)\begin{pmatrix}\cos\psi_i\\ \sin\psi_i\end{pmatrix}$$

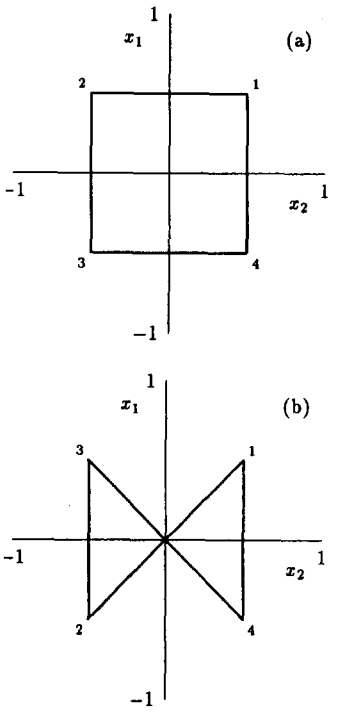

Fig. 3. Configurations in a two-dimensional map: **a** rectangle, **b** twist. See the text for further explanation

The radius $r(\sigma)$ is chosen such that the global minima lie on this circle. We obtain

$$r(\sigma) = \frac{1}{\sqrt{2}}\frac{1-\sigma}{1+\sigma}$$

2. The first weight vector is fixed to cover the first quadrant:

$$\psi_1 = \pi/4$$

3. The sum of all angles ψ_i is constant:

$$\sum_{i=1}^{4} \psi_i = 4\pi$$

With these constraints the weight vector is fully described by two parameters ω_1 and ω_2, defined by

$$\omega_1 \stackrel{\text{def}}{=} (\psi_4 - \psi_2)/\pi, \quad \omega_2 \stackrel{\text{def}}{=} (2\psi_3 - \psi_2 - \psi_4)/\pi$$

If $\vec{w}_1$ covers the first quadrant, there are still six different ways to cover the other three quadrants. Therefore there are six different minima: two rectangles and four twists. In terms of ω_1 and ω_2 they are positioned as follows:

Minimum	rectangle	twist	twist	rectangle	twist	twist
Configuration	(1234)	(1243)	(1423)	(1432)	(1342)	(1324)
(ω_1, ω_2)	(1, 0)	(1/2, 1)	(−1, 0)	(−1/2, 1)	(−1/2, −1)	(−1/2, −1)

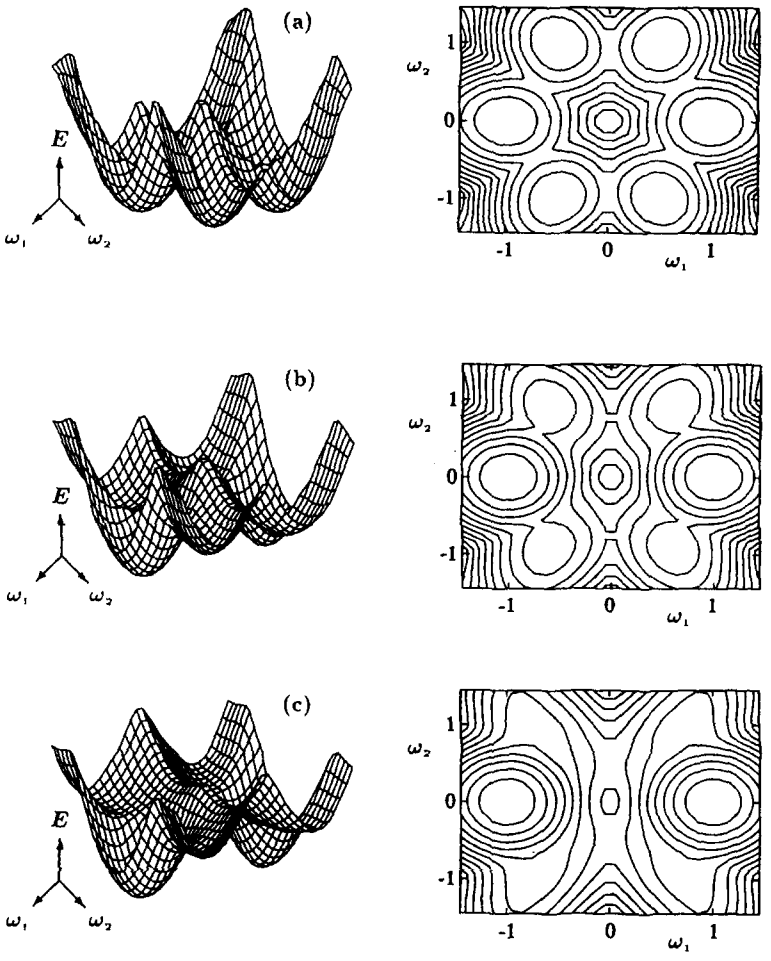

Fig. 4. The error potential E as a function of the parameters ω_1 and ω_2 for different values of the interaction strength σ: **a** $\sigma = 0$, **b** $\sigma = 0.15$, **c** $\sigma = 0.3$. Contour plots are shown on the *right*. See the text for further explanation

The error potential as a function of the parameters ω_1 and ω_2 is plotted in Fig. 4 for three different values of σ. Again, all minima are equally deep for $\sigma = 0$ (Fig. 4a) and this symmetry is broken for $\sigma > 0$ (Fig. 4b). If we raise the interaction strength σ the local minima eventually disappear (Fig. 4c).

5 Transition times

Because of the random presentation of the input vectors and (possibly) the random initialization of the weights, the learning process is a stochastic process governed by the master equation (Ritter and Schulten 1988; Heskes and Kappen 1991)

$$\frac{\partial P(\mathbf{w}', t)}{\partial t} = \int d^N w [T(\mathbf{w}' \mid \mathbf{w}) P(\mathbf{w}, t) - T(\mathbf{w} \mid \mathbf{w}') P(\mathbf{w}', t)] \quad (6)$$

with $P(\mathbf{w}, t)$ the probability that the network is in state $\mathbf{w}$ at time t. The transition probability $T(\mathbf{w}'|\mathbf{w})$ is the probability of drawing an input vector $\vec{x}$ such that the learning rule (5) changes the weight vector from $\mathbf{w}$ to $\mathbf{w}'$:

$$T(\mathbf{w}' \mid \mathbf{w}) = \int d^m x \, \rho(\vec{x}) \delta^N(\mathbf{w}' - \mathbf{w} - \eta \mathbf{f}(\mathbf{w}, \vec{x}))$$

Ritter and Schulten (1988) used a Fokker-Planck approach to approximate this master equation and to study the final convergence to a global minimum. However, this Fokker-Planck approach fails to describe global properties of the learning process such as transition times between different minima. To make some progress we will have to make a few assumptions and approximations. Details can be found in Heskes et al. (1992).

Let us divide the weight space into attraction regions and transition regions. In attraction regions the Hessian matrix $H(\mathbf{w})$ with components

$$H_{i\alpha j\beta}(\mathbf{w}) \stackrel{\text{def}}{=} \frac{\partial^2 E(\mathbf{w})}{\partial w_{i\alpha} \partial w_{j\beta}}$$

is a positive definite matrix, i.e., in the attraction regions all eigenvalues of the Hessian $H(\mathbf{w})$ are positive, whereas in the transition regions at least one of the eigenvalues is negative. Each attraction region $\mathcal{K}$ contains one minimum $\mathbf{w}_k^*$ of the error potential $E(\mathbf{w})$.

On a time scale of order $1/\eta$, the probability density function $P(\mathbf{w}, t)$ becomes a distribution with peaks in the attraction regions. We expand the probability density function $P(\mathbf{w}, t)$ as a sum over the functions $P_k(\mathbf{w}, t)$ in the attraction regions and $P_{\text{rest}}(\mathbf{w})$ in the transition regions:

$$P(\mathbf{w}, t) = \sum_k P_k(\mathbf{w}, t) + P_{\text{rest}}(\mathbf{w}, t)$$

By definition $P_k(\mathbf{w}, t)$ is zero outside attraction region $\mathscr{K}$ and $P_{rest}(\mathbf{w}, t)$ is zero outside the transition regions. For small learning parameters the probability mass in the transition regions is negligible in comparison with the probability mass in the attraction regions. The first assumption now states that in the attraction regions time and space decouple, i.e., that we may write

$$P_k(\mathbf{w}, t) = n_k(t)\, p_k(\mathbf{w})$$

with $p_k(\mathbf{w})$ a local normalized distribution and $n_k(t)$ the occupation number in attraction region $\mathscr{K}$. The underlying assumption is here that the interaction between different minima may affect the total probability *mass* in the attraction region $\mathscr{K}$, indicated by the occupation number $n_k(t)$, but not the *shape* of the local distribution, indicated by $p_k(\mathbf{w})$. This assumption is frequently used in the theory on unstable stochastic processes and seems to be valid if the attraction regions are well separated and transitions between them are rare.

According to Van Kampen's expansion (Van Kampen 1981), the asymptotic expansion of the distribution $p_k(\mathbf{w})$ for small learning parameters η is a Gaussian, denoted by $G_k(\mathbf{w})$. The first moment of this Gaussian is the minimum $\mathbf{w}_k^*$. Its covariance matrix Σ_k^2 can be written

$$\Sigma_k^2 = \eta K_k \tag{7}$$

with K_k independent of the learning parameter η. K_k is the solution of the matrix equation

$$HK_k + K_k H = D \tag{8}$$

where the elements of the Hessian H and diffusion matrix D are given by $H_{i\alpha j\beta} \equiv H_{i\alpha j\beta}(\mathbf{w}_k^*)$ and $D_{i\alpha j\beta} \equiv \langle f_{i\alpha}(\mathbf{w}_k^*, \vec{x}) f_{j\beta}(\mathbf{w}_k^*, \vec{x})\rangle_\Omega$. The Hessian is related to the curvature of the error potential, the diffusion to the fluctuations in the learning rule; both are evaluated at the minimum $\mathbf{w}_k^*$.

Let us try to calculate the transition time τ_{lk} from an attraction region $\mathscr{K}$ to another attraction region $\mathscr{L}$. We restrict ourselves to transitions such that attraction region $\mathscr{L}$ can be reached from $\mathscr{K}$ through an intermediate transition region $\mathscr{T}$ in which the Hessian $H(\mathbf{w})$ has only one negative eigenvalue. Transitions from kinks to lines and from twists to rectangles fulfill this requirement. The second assumption now claims that the dominant contribution to this transition time stems from the transition probability Γ_{tk} of going from attraction region $\mathscr{K}$ to transition region $\mathscr{T}$. Theoretical arguments and experimental checks can be found in Heskes et al. (1992). Basically, the idea is that in the attraction regions the variance of the local fluctuations converges to a constant proportional to the learning parameter [see (7)], whereas in the transition regions this variance shows a tendency to diverge, i.e., is hardly affected by the choice of the learning parameter. The transition probability from attraction region $\mathscr{K}$ to transition region $\mathscr{T}$ reads

$$\Gamma_{tk} = \int_{\mathscr{T}} d^N w' \int_{\mathscr{K}} d^N w\, T(\mathbf{w}' \mid \mathbf{w})\, G_k(\mathbf{w}) \tag{9}$$

where $G_k(\mathbf{w})$ is the Gaussian given by Van Kampen's expansion. In (9), we have to integrate over all $\mathbf{w}$ and $\vec{x}$ such that

$$\mathbf{w} \in \mathscr{K}$$

and

$$\mathbf{w}' = \mathbf{w} + \eta \mathbf{f}(\mathbf{w}, \vec{x}) \in \mathscr{T}$$

So, both $\mathbf{w}$ and $\mathbf{w}'$ are within order η of the boundary $\mathscr{T}\mathscr{K}$ between attraction region $\mathscr{K}$ and transition region $\mathscr{T}$. Now it is easy to prove (Heskes et al. 1992) that, for small learning parameters η, the integral in (9) converges to an integral over the boundary $\mathscr{T}\mathscr{K}$ times some term of order η. In the limit $\eta \to 0$ this integral over the boundary $\mathscr{T}\mathscr{K}$ can be computed using the method of steepest descent: the largest contribution is found for the weight vectors $\mathbf{w}$ with the largest $G_k(\mathbf{w})$ on the boundary $\mathscr{T}\mathscr{K}$. In other words, the most important contribution to the transition time τ_{lk} stems from the 'easiest' path from the local minimum $\mathbf{w}_k^*$ to the transition region $\mathscr{T}$. The matrix K_k^{-1} defines the local 'metric'. Our final result is

$$\mathscr{T}_{lk} \approx \frac{1}{\Gamma_{tk}} \sim \exp\left[\frac{\tilde{\eta}_{lk}}{\eta}\right], \quad \text{for } \eta \to 0 \tag{10}$$

with the so-called reference learning parameter

$$\tilde{\eta}_{lk} = \inf_{\mathbf{w} \in \mathscr{T}\mathscr{K}} (\mathbf{w} - \mathbf{w}_k^*)^T K_k^{-1} (\mathbf{w} - \mathbf{w}_k^*)/2 \tag{11}$$

Roughly speaking, the reference learning parameter is proportional to the height of the error barrier and inversely proportional to the fluctuations in the learning rule. It is similar to the Arrhenius factor in chemical reaction theory (e.g., Van Kampen 1981). A reasonable estimate for the reference learning parameter is desirable, since for $\eta \ll \tilde{\eta}_{lk}$ the probability of going from minimum $\mathbf{w}_k^*$ to minimum $\mathbf{w}_l^*$ within an acceptable number of learning steps is negligible. Furthermore, the reference learning parameter is the key parameter in cooling schedules for the learning parameter that guarantee convergence to the global minimum (Heskes et al. 1993).

We summarize the most important conclusions of this section. The transition times between different minima grow exponentially with the quotient of the reference learning parameter $\tilde{\eta}$ and the learning parameter η. We think that we know how to calculate this reference learning parameter. However, we must be careful, since our calculation is based on two hypotheses:

1. The transitions between various minima may affect the mass, but not the (Gaussian) shape of the local distributions in the attraction regions.
2. To calculate, or at least estimate, the reference learning parameter for the transition from attraction region $\mathscr{K}$ to attraction region $\mathscr{L}$ it is sufficient to compute the reference learning parameter for the transition from attraction region $\mathscr{K}$ to the intermediate transition region $\mathscr{T}$.

6 Theory versus simulations

It this section we will compare the reference learning parameters predicted by the theory given in Sect. 5 with the reference learning parameters obtained from simulations of the learning process. We will study both the kinks in one-dimensional maps (Sect. 3) and the twists in two-dimensional maps (Sect. 4). We will focus on the transition time from a disordered local minimum [the kink (132) and the twist (1324)] to the perfectly ordered global minimum [the line (123) and the rectangle (1234)]. Transitions from perfectly ordered configurations to topological defects are far less probable, except for a small lateral-interaction strength σ and a relatively large learning parameter η. Note that this is in contrast to the original Kohonen learning rule where it is *impossible* instead of very *improbable* to leave a perfectly ordered one-dimensional mapping (Kohonen 1988). This is particular for one-dimensional mappings: there are no ordering proofs for higher-dimensional mappings.

To calculate the reference learning parameters predicted by theory we go through the following steps:

1. Choose the lateral-interaction strength σ.
2. Determine the position of the local minimum $\mathbf{w}_k^*$.
3. Calculate the Hessian H and the diffusion matrix D at this minimum.
4. Solve (8) to find the covariance matrix K_k and its inverse K_k^{-1}.
5. Find the point $\mathbf{w}$ on the boundary between the attraction and the transition region, i.e., where the determinant of the Hessian of the error potential $E(\mathbf{w})$ is exactly zero, with the smallest distance $(\mathbf{w} - \mathbf{w}_k^*)^T K_k^{-1} (\mathbf{w} - \mathbf{w}_k^*)$.
6. Compute the reference learning parameter using (11).

The calculation of the determinant of the Hessian matrix for general $\mathbf{w}$ makes the fifth step the most difficult one. The Hessian matrix for the twists is calculated numerically, using the error potential (3) not in the limit $\beta \to \infty$, but for large $\beta = 50$. This results in a tiny error, negligible in comparison with the numerical precision of the simulations. The continuous lines in Fig. 5a and b give the theoretical reference learning parameters for the kinks and twists, respectively.

Straightforward simulations of the learning rule (5) will be used for comparison. For every choice of the lateral-interaction strength σ, we train 500 independently operating networks for four different learning parameters. Each simulation starts with the networks near a local minimum. The transition time is measured from the exponential decay of the occupation number at the local minimum. Theory and simulations predict transition times $\tau(\eta)$ of the form (Heskes et al. 1992)

$$\ln \tau(\eta) = \tilde{\eta}/\eta - d \ln \eta + c \qquad (12)$$

If possible, the learning parameters for each value of the interaction strength σ are chosen such that this transition time varies from about 10^2 learning steps for the largest learning parameter to about 10^5 learning steps for the smallest learning parameter. This is intractable for small σ, since for stability of the learning rule (5) we must keep $\eta < 1$, whereas for $\eta \ll 1$ the simulations become too time-consuming. So, for small σ, we cannot simulate four totally different learning parameters, which makes it difficult to obtain accurate estimates for both $\tilde{\eta}$ and d in (12). However, looking at simulations for larger σ, it seems that the parameter d is near 0.5, independent of the lateral-interaction strength σ. So, we keep d fixed at $d = 1/2$, which would also follow from a more detailed derivation of the transition time (10) under the assumption that the local probability shape is a perfect Gaussian (assumption 1) and that the influence of the transition region on the total transition time can be neglected completely (assumption 2) (see also Heskes et al. 1992). The parameters $\tilde{\eta}$ and c are now calculated from a least square fit of formula (12) through the four points. The numerical precision of these least square fits (one fit for each value of the interaction strength σ) is about 10% of the result. The reference learning parameters $\tilde{\eta}$ obtained in this way are indicated by an asterisk in Fig. 5.

Qualitatively, we obtain good agreement between theory and simulations, except for small values of σ in the one-dimensional map. The 'experimental' reference learning parameter seems to diverge to infinity for $\sigma \to 0$, whereas the 'theoretical' reference learning parameter tends to 9/2. The limit $\sigma \to 0$ is quite peculiar, since for $\sigma = 0$ ergodicity is broken: it is impossible to change the ordering of the weights for $\eta < 1$ and thus the reference learning parameter for this transition is indeed infinite. In fact, this is also predicted by theory. Since it is impossible to cross the boundary $\mathcal{TK}$, the transition probability Γ_{tk}, as defined in (9), is zero.

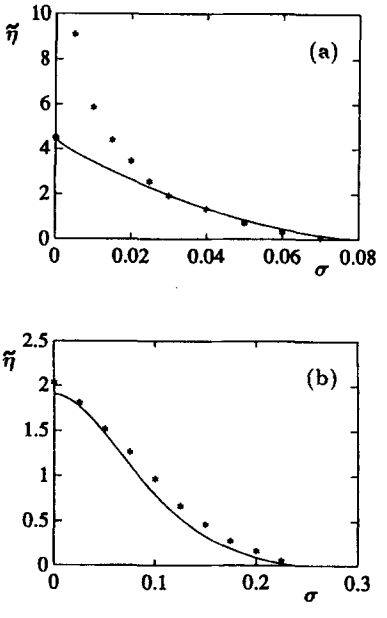

Fig. 5. The reference learning parameter $\tilde{\eta}$ as a function of the interaction strength σ for kinks (**a**) and twists (**b**). *Continuous lines* show the theoretical results. Simulation results are indicated by *asterisks*

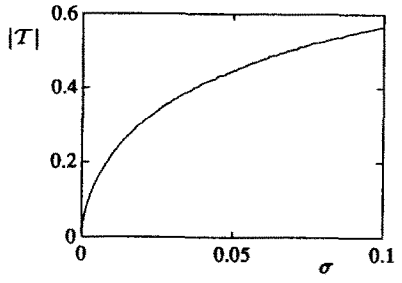

Fig. 6. The relative volume $|\mathscr{T}|$ of the transition region as a function of the interaction strength σ

For nonzero σ and any nonzero learning parameter η, there exists a sequence of inputs x such that w_2 and w_1 interchange and thus there is a nonzero transition probability from the local minimum with configuration (213) to the global minimum (123). Why does the theory fail to give the right results here? Apparently, the hypotheses we had to make to calculate the reference learning parameter are no longer valid in this limit.

To check whether we really can neglect the influence of the transition region on the total transition time from one attraction region to another, we calculate the relative volume of the transition region as a function of σ. We take a box $[1/6, 5/6] \times [1/6, 5/6] \times [1/6, 5/6]$. This box just includes all minima. Outside this box the Hessian matrix is positive definite everywhere. The relative volume $|\mathscr{T}|$ of the transition region is defined as the fraction of this box in which at least one eigenvalue $\lambda_i(\mathbf{w})$ of the Hessian matrix $H(\mathbf{w})$ is negative:

$$|\mathscr{T}| \stackrel{\text{def}}{=} \frac{\int_{\text{box}} d^3w [1 - \prod_{i=1}^{3} \theta(-\lambda_i(\mathbf{w}))]}{\int_{\text{box}} d^3w}$$

Figure 6 shows the results obtained through stochastic integration using 10^6 points. We conclude that the volume of the transition region goes to zero for $\sigma \to 0$. Looking at the second assumption, we are tempted to say that this should make the difference between simulations and theory smaller and certainly not larger.

Yet, the gap between theory and simulations can be explained by looking at the volume of the transition region. A very small transition region may help to defend the second assumption; it is the death-blow for the first assumption. In our calculation of the reference learning parameter we need to know the shape of the local probability density function at the boundary between the attraction region and the transition region. The assumption is that this shape depends only on local properties of the learning rule (the curvature H and the diffusion D) at the local minimum, and not on transitions between the two attraction regions, i.e., not on properties of the learning rule in the other attraction region. The transition region acts like a buffer. It separates the local distributions in the two attraction regions, i.e., it allows for a proper linking of the two distributions $n_k(t)p_k(\mathbf{w})$ and $n_l(t)p_l(\mathbf{w})$ in the attraction regions. However, in the case of a very small transition region between two attraction regions, the buffer becomes too small: it is no longer justified to treat the shapes of the local probability distributions in the attraction regions as though they are independent, i.e., the continuous probability distribution $P(\mathbf{w}, t)$, the true solution of the master equation (6), cannot be approximated by two independent parts $n_k(t)p_k(\mathbf{w})$ and $n_l(t)p_l(\mathbf{w})$. The first assumption is therefore violated and the theoretical results obtained are meaningless. Alas, this assumption is crucial in our analysis: we do not know how to calculate or estimate the reference learning parameter if it is no longer true.

A full Fokker-Planck description of the learning process can be considered as an alternative to the approach taken in this paper. In Heskes (1994) it has been shown that these two approaches are more or less equivalent in the sense that they yield qualitatively similar and reasonable results. Both being invalid approximations inside the transition regions, neither of the two approaches can be used for exact computation of the reference learning parameter.

7 Discussion

In this paper we defined an error potential for the self-organization of topological maps. The corresponding learning rule is the original Kohonen learning rule, except for the determination of the winning unit. Our learning rule is computationally more expensive. The Euclidian distances, which require $n \times m$ multiplications (n is the number of topological units, m the dimension of the input vectors), must be multiplied with the interaction matrix h. This requires $n \times |h|$ extra multiplications, with $|h|$ the number of nonzero lateral connections for one unit. The determination of the winning unit requires n operations. So, our learning rule is about a factor $(m + |h| + 1)/(m + 1)$ slower than the original Kohonen learning rule. This is the price we have to pay for knowing exactly what is minimized by the learning procedure. So, from a *theoretical* point of view, a learning rule derived from an error potential is more elegant and often easier to analyze. For *practical* applications the closely related Kohonen learning rule is faster and therefore favorable.

In our analysis, both the interaction strength σ and the learning parameter η were kept constant. We found that the transition times between different minima can be written (for small learning parameters η)

$$\tau(\eta, \sigma) \sim \exp\left[\frac{\tilde{\eta}(\sigma)}{\eta}\right]$$

where the reference learning parameter $\tilde{\eta}(\sigma)$ is a function of the interaction strength σ. Changing the interaction strength σ means changing the error potential as a function of time. For slow changes in σ (slow on a time scale of one over the learning parameter, the typical 'local' convergence time), the probability distribution in the attraction region can still be approximated by a Gaussian with covariance matrix determined by the curvature of the error potential and the fluctuations in the learning

rule at the minimum. A similar 'quasi-stationary' argument holds for a gradually changing learning parameter (Heskes et al. 1993). So, for slow changes we may approximate

$$\tau(\eta(t), \sigma(t)) \sim \exp\left[\frac{\tilde{\eta}(\sigma(t))}{\eta(t)}\right]$$

Of course, the stronger the time dependency of either η or σ, the less accurate this quasi-stationary approximation.

The error potential (3) is well defined for *any* lateral-interaction matrix h, symmetric or asymmetric, subject to normalization or free of constraints. The symmetry and normalization constraints used in the examples are just *choices* to simplify the analysis. With the normalization constraint (1) the 'receptive fields' (parts of the input space where a unit is the 'winner') are bounded by linear manifolds; without this constraint the separatrices between receptive fields may be curved. At first sight, symmetry in the lateral-interaction matrix is the most obvious choice and leaves us with only one adjustable parameter (besides the learning parameter): the lateral-interaction strength σ. However, asymmetry in the lateral-interaction matrix introduces a certain bias and may therefore lead to faster ordering (Geszti 1990). Because of its generality, the error potential (3) can be used to study this claim.

We kept the examples in this paper as small as possible: kinks in a one-dimensional map consisting of three units and twists in a two-dimensional map consisting of four units. In this way we were able to make pictures of the error potential and to compare theoretical results with simulations. Kinks in one-dimensional maps and twists in two-dimensional maps are the most appealing examples of local minima in self-organizing maps. However, in practical applications self-organizing maps are most often used to map a higher-dimensional input space onto a lower-dimensional network structure. Therefore, it seems worthwhile to study local minima and transition times for toy problems as in Ritter and Schulten (1988), where a three-dimensional input space is mapped on a two-dimensional network structure. In general, we believe that the principles emerging in this study will, at least *qualitatively*, also apply to larger problems, but that it will require a considerable amount of computational effort and theoretical study to obtain *quantitative* results on the global performance of large self-organizing networks.

Acknowledgements. This work was partly supported by the Dutch Foundation for Neural Networks. I would like to thank Andrzej Komoda, Eddy Slijpen and Stan Gielen for critically reading a previous version of this manuscript. I am also grateful to Prof. Teuvo Kohonen for some useful comments.

References

Bauer H-U, Pawelzik KR (1992) Quantifying the neighborhood preservation of self-organizing feature maps. IEEE Trans Neural Networks 3:570–579

Cottrell M, Fort JC (1987) A stochastic model of retinotopy: a self-organizing process. Biol Cybern 53:405–411

Durbin R, Mitchison G (1990) A dimension reduction framework for understanding cortical maps. Nature 343:644–647

Erwin E, Obermayer K, Schulten K (1992) Self-organizing maps: ordering, convergence properties and energy functions. Biol Cybern 67:47–55

Geszti T (1990) Physical models of neural networks. World Scientific, Singapore

Heskes TM (1994) On Fokker-Planck approximations of on-line learning processes. J Phys A 27:5145–5160

Heskes TM, Kappen B (1991) Learning processes in neural networks. Phys Rev A 44:2718–2726

Heskes TM, Kappen B (1993) Error potentials for self-organization. In: International Conference on Neural Networks, San Francisco, vol III. IEEE, New York, pp 1219–1223

Heskes TM, Slijpen ETP, Kappen B (1992) Learning in neural networks with local minima. Phys Rev A 46:5221–5231

Heskes TM, Slijpen ETP, Kappen B (1993) Cooling schedules for learning in neural networks. Phys Rev E 47:4457–4464

Kohonen T (1982) Self-organized formation of topologically correct feature maps. Biol Cybern 43:59–69

Kohonen T (1988) Self-organization and associative memory. Springer, Berlin Heidelberg New York

Kohonen T (1991) Self-organizing maps: optimization approaches. In: Kohonen T et al. (eds) Artificial neural networks, vol II. North-Holland, Amsterdam, pp 981–990

Luttrell SP (1989) Self-organisation: a derivation from first principles of a class of learning algorithms. In: International Joint Conference on Neural Networks, vol II. IEEE Computer Society Press, pp 495–498

Luttrell SP (1994) A Bayesian analysis of self-organizing maps. Neural Comput 6:767–794

Miller K, Keller J, Stryker M (1989) Ocular dominance column development: analysis and simulation. Science 245:605–615

Obermayer K, Ritter H, Schulten K (1990) A principle for the formation of the spatial structure of cortical feature maps. Proc Natl Acad Sci USA 87:8345–8349

Obermayer K, Blasdel G, Schulten K (1992) Statistical-mechanical analysis of self-organization and pattern formation during the development of visual maps. Phys Rev A 45:7568–7589

Ritter H, Schulten K (1986) On the stationary state of Kohonen's self-organizing sensory mapping. Biol Cybern 54:99–106

Ritter H, Schulten K (1988) Convergence properties of Kohonen's topology conserving maps: fluctuations, stability, and dimension selection. Biol Cybern 60:59–71

Ritter H, Obermayer K, Schulten K, Rubner J (1991) Self-organizing maps and adaptive filters. In: Domany E et al. (eds) Models of neural networks. Springer, Berlin Heidelberg New York, pp 281–306

Rose K, Gurewitz E, Fox GC (1990) Statistical mechanics of phase transitions in clustering. Phys Rev Lett 65:945–948

Takeuchi A, Amari S (1979) Formation of topographic maps and columnar microstructures. Biol Cybern 35:63–72

Tolat VV (1990) An analysis of Kohonen's self-organizing maps using a system of energy functions. Biol Cybern 64:155–164

Van Kampen NG (1981) Stochastic processes in physics and chemistry. North-Holland, Amsterdam

Von der Malsburg Ch (1973) Self-organization of orientation sensitive cells in the striate cortex. Kybernetik 14:85–100